Barycentric Defense

A V Subramanyam  
Department of Electronics and Communication Engineering  
Indraprastha Institute of Information Technology  
New Delhi, India  
subramanyam@iiitd.ac.in

Abhigyan Raj  
Department of Computer Science  
Indian Institute of Technology  
Dhanbad, India  
abhigyan.18je0013@cse.iitism.ac.in

Abstract—Wasserstein metric based adversarial attacks have attracted a great interest in the recent past. Even though they exhibit strong attacks, surprisingly, they have not been investigated for defense. In this work, we demonstrate that barycenters computed in Wasserstein space can act as a measure of defense against adversarial attacks. We compute the barycenter using marginals obtained from the given image and demonstrate its effectiveness in defense even without any adversarial training. We further analyse the barycenters using GradCam to understand their defensive characteristics. Elaborate experiments on MNIST, Fashion-MNIST, CIFAR-10 and CIFAR-100 demonstrate a significant increase in the robustness of victim classifiers.

Index Terms—Barycenter, Dual Wasserstein, Adversarial defense.

I. INTRODUCTION

Deep learning systems have shown impressive performance in various applications. However, these systems are vulnerable to adversarial perturbations [1]–[3]. In order to counter these attacks, several defense mechanisms have also been proposed.

In one of the early works, Szegedy et al. [4] formulated the adversarial attack as an optimization problem and obtained the adversarial sample using L-BFGS. Several adversarial attacks have been proposed since Szegedy’ work [5], [6]. In particular, a vast majority of attacks are proposed using $l_p$ norm only. On the other hand, strong defense measures have been studied in [1], [7]–[9].

Different from $l_p$ norm based attacks, Wasserstein distance based attacks have also been studied in [10], [11]. Unlike pixel-wise distances in $l_p$ space, Wasserstein distance incorporates the geometry of the vectors between which the distance is computed. This makes the Wasserstein space more attractive as minor transformations do not lead to large distances in contrast to other metrics such as Euclidean distance. However, defense based on Wasserstein distance is not yet explored and this is the focus of our work.

In this work, we propose a barycenter based defense in a classification setting. Our hypothesis is that the attacks designed in $l_p$ space may not be effective in the Wasserstein space. In order to verify this, we find the barycenter of the given image using marginals derived from the given image itself. We find that the barycenter is very effective in defending against $L_{inf}$ and $L_2$ attacks. We also investigate the gradcam [12] images where we demonstrate that the barycenter actually possesses traits of the original image itself which is the reason for effectiveness of barycenter against attacks.

II. RELATED WORKS

Adversarial Attacks Some of the robust attacks are iterative FGSM (Kurakin et al. [6]), PGD (Madry et al. [7]), Carlini and Wagner [13] attacks, Jacobian based attack (Papernot et al. [14]), physical attack (Athalye et al. [15]), Autoattack (Croce and Hein [2]), momentum based iterative FGSM (Dong et al. [16]), single pixel attack (Su et al. [17]). These attacks are primarily focused in $l_p$ domain. Attacks in Wasserstein space have also been explored in [10], [11], [18], [19].

Adversarial Defense In response to adversarial attacks, adversarial defense has been proposed to defend the victim models. One of the best defense approach is adversarial training [4], [5], [20]. Madry et al. [7] formally studied adversarial training and proposed that such training allows network to defend well against first order adversary. Adversarial logit pairing uses a pair of logits from clean and adversarial examples to defend against adversarial samples [21]. Prominent theoretical studies include TRADES [22] which prove the bounds based on additional regularization term that minimizes the difference in prediction between clean and adversarial samples. In [1], Wong et al. proposed to effectively combine FGSM and random initialization to
where, \( \eta \) denotes inner product.

**Wasserstein Barycenter** In the following we discuss Wasserstein distance and barycenter. The entropic regularized Wasserstein distance is given as [33],

\[
W_{\eta}(x, y) = \min_{\Pi} (\Pi, C) + \eta(\Pi, \ln \Pi)
\]

s.t. \( \Pi x = x, \Pi^T 1 = y \),

where, \( x \) and \( y \) denote probability simplexes, \( \Pi \) denotes the transport plan, \( C \) denotes the cost matrix, \( 1 \) denotes a vector of ones, \( \eta \) is a regularization parameter, and \( \langle \cdot, \cdot \rangle \) denotes inner product.

Let there be \( N \) marginals \( q_k, k = 1, 2, …, N \), and barycenter weights \( \beta_k \). Then the barycenter \( p \) is the solution of the following objective [34],

\[
\min_p \sum_{k=1}^{N} \beta_k W_{\eta}(p, q_k).
\]

The regularized version of the barycenter is given as,

\[
\min_p \sum_{k=1}^{N} \beta_k W_{\eta}(p, q_k) + J(p)
\]

where \( J(p) \) can be an \( l_2 \) regularizer, for example, \( J(p) = \frac{\lambda}{2} \|p\|_2^2 \).

**III. METHODOLOGY**

We discuss our proposed objective formulation in this section. We first formulate a regularized barycenter problem and derive a solution to obtain the barycenter. Let the clean or adversarial sample be \( x \). Let \( \mathcal{A} \) be a linear operator, then we define the marginal of \( x \) as \( q_k = \mathcal{A}_k x \). In our experiments, we use rotation and translation as the linear operator. Then the regularized Wasserstein barycenter is given as,

\[
\min_p \sum_{k=1}^{N} \beta_k W_{\eta}(p, q_k) + J(p),
\]

where \( J(p) = 0.5\theta \|x - p\|_2^2 \).

We choose a \( l_2 \) regularizer different from the regularizer in [34]. This is because barycenter computation can be unstable when a regularizer is only focusing on \( p \). In our objective function Equation 1, we try that the barycenter is also close to the given sample \( x \) in the \( l_2 \) space and thus we use \( J(p) = 0.5\theta \|x - p\|_2^2 \).

In order to obtain \( p \), we solve the dual problem of Equation 1 given by,

\[
\min_{(u_k)_{k=1}^N, v} \sum_{k=1}^{N} \beta_k H^*(q_k) + J^*(v)
\]

where \( H^*(q_k) = \gamma [E(q_k) + \langle q_k, \ln K \alpha_k \rangle] \) and \( E(q_k) = -\sum_i q_k^i \alpha_k = e^{u_k} / \gamma \).

We can expand \( J^*(v) \) as,

\[
J^*(v) = \sup_s (v^T s - J(s)) = -\|v\|_2^2 / 2\gamma + v^T w.
\]

The solution for Equation 2 is given in Algorithm 1.

**Algorithm 1: Regularized Barycenter**

**Input:** \((u_k, v) \leftarrow 1 \) a vector of ones with dimension \( n \times 1 \), \( N \) marginals \( q_k \),

\( k = 1, 2, …, N \); \( \beta_k = 1 / N \) be the barycenter weights, \( K = e^{-C / \gamma} \), \( \gamma = 0.02 \), \( \theta = 1e5 \), \( \tau = 0.2 \)

**Output:** \( p = \frac{1}{N} \sum_k \nabla H^*_\eta(q_k) \)

**while max iterations do**

\[
\begin{align*}
&u_N \leftarrow -\frac{u_N}{\beta_N} - \sum_{k=1}^{N-1} \beta_k u_k \\
&\nabla H^*_\eta(q_k) = \alpha_k \odot K q_k \\
&\nabla F((u_k)_{k=1}^N, v) = \left[ (\beta_k \nabla H^*_\eta(q_k)) - \nabla H^*_\eta(v) \right]_{k=1}^{N-1} - \nabla H^*_\eta(u_N) \\
&x^{l+1} = \text{Prox}_{\tau J^*}(x^l - \tau \nabla F(x^l)) \\
&x^{l+1} = x^{l+1} - \tau \nabla F(x^{l+1})
\end{align*}
\]

**end**

Here, \( \text{Prox}_{\tau J^*}(v) = \arg \min_{v'} 0.5\|v - v'\|_2^2 + \tau J^*(v') \).

**IV. EXPERIMENTAL RESULTS**

**A. Implementation Details**

The proposed method is implemented in Pytorch framework. In case of MNIST and Fashion-MNIST, we use the architecture proposed in [10]. For CIFAR-10, we use the Resnet-18 pre-trained model available at [35]. MNIST model obtains an accuracy of 99.0% on clean examples, Fashion-MNIST model achieves an accuracy of 90.0%, and CIFAR-10 model obtains an accuracy...
of 94.8% on clean examples. For CIFAR-100, we use Resnet-32 with clean sample accuracy of 70.14%. We use the local cost matrix [10] obtained using \( l_1 \) norm between \((i, j)\) and \((k, l)\) as \(|i-j|+|k-l|\) co-ordinates. The size of the local cost matrix is 7x7.

### B. Adversarial attack performance

In Table I, we show the performance of clean samples, barycenters and adversarial samples obtained using \( L_\infty \) attacks. We compute the barycenters in the following two settings. First, we use 6 marginals comprising of the given image (attacked or clean), \( \pm 2^\circ \) rotation of the given image, \( \pm 3^\circ \) rotation of the given image and translation by 3 pixels. We refer this setting as \( B_6 \) in Table I. In the second setting, we use the given image (attacked or clean), \( \pm 2^\circ \) rotation of the given image, and translation by 3 pixels. We refer this as \( B_4 \). We tried with higher degrees of rotation, however, too much rotation itself degrades the model performance. Therefore, a rotation of utmost \( \pm 3^\circ \) is considered. The choice of translation is also motivated by common data augmentation methods.

#### Table I

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( L_\infty )</th>
<th>Barycenter</th>
<th>Clean</th>
<th>Barycenter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( B_6 )</td>
<td></td>
<td>( B_4 )</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>0</td>
<td>53.8</td>
<td>94.8</td>
<td>88.4</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>90.3</td>
<td>99.0</td>
<td>98.9</td>
</tr>
<tr>
<td>MNIST</td>
<td>0.5</td>
<td>67.1</td>
<td>90.0</td>
<td>89.8</td>
</tr>
<tr>
<td>F-MNIST</td>
<td>0.5</td>
<td>27.2</td>
<td>70.2</td>
<td>60.5</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>0</td>
<td>21.2</td>
<td>70.2</td>
<td>64.6</td>
</tr>
</tbody>
</table>

It can be observed that the accuracy under \( L_\infty \) is almost zero. Whereas, in case of barycenters, the accuracy is very high. A similar trend is observed for all datasets as well as for both attacked and clean images. Thus, barycenters can achieve an accuracy close to that of victim model in case of clean images, and also defend the model under adversarial attack.

We show the performance against several attacks in Tables II and III. We use Foolbox [36], [37] to perform the attacks. We report the results in terms of attack accuracy (AA) and barycenter accuracy (BA). AA is the ratio of correctly classified attacked samples to that of the total number of samples. It is computed over the complete test set. We can observe that AA is very low when the adversarial samples are presented to the victim model. On the other hand, when we test using the barycenters generated from these attacked images, the accuracy is high. We denote this accuracy as Barycentric Accuracy (BA).

#### Table II

<table>
<thead>
<tr>
<th>Attack</th>
<th>AA (in %)</th>
<th>BA (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2PGD [7]</td>
<td>43.59</td>
<td>62.03</td>
</tr>
<tr>
<td>L2Carliniwagner [13]</td>
<td>63.19</td>
<td>96.07</td>
</tr>
<tr>
<td>Linfpgd [7]</td>
<td>16.36</td>
<td>31.75</td>
</tr>
<tr>
<td>FGSM [5]</td>
<td>71.75</td>
<td>74.96</td>
</tr>
<tr>
<td>Linfdeepfool [20]</td>
<td>42.52</td>
<td>94.69</td>
</tr>
</tbody>
</table>

#### Table III

<table>
<thead>
<tr>
<th>Attack</th>
<th>AA (in %)</th>
<th>BA (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGSM [5]</td>
<td>45.88</td>
<td>51.6</td>
</tr>
<tr>
<td>Linfdeepfool [20]</td>
<td>42.59</td>
<td>81.05</td>
</tr>
<tr>
<td>AutoAttack [2]</td>
<td>0.6</td>
<td>2.75</td>
</tr>
</tbody>
</table>

We present the results for CIFAR-10 and CIFAR-100 in Tables IV and V, respectively. In both these cases, we find that our model exhibits a strong defense against attacks like L2Carliniwagner [13] and Linfdeepfool [20]. Further, as AutoAttack [2] is quite powerful, it is hard to achieve a sound robustness against this attack without doing any adversarial training. We also performed experiments by increasing \( \epsilon \) to 0.3 for MNIST and 16/255 for CIFAR-10, however, both attack and barycenter accuracies are low in this case.

Although methods involving adversarial training such as [7], [22] achieve a high accuracy compared to that of ours, the proposed method opens up a new direction for exploring defense without any training.

#### C. Sensitivity analysis

We show the sensitivity analysis in Figure 1. The image enclosed in red box denotes the barycenter obtained from the settings used for all the experiments. We fix \( \theta \) and \( \tau \), and increase \( \gamma \) to analyse the effect on barycenter. This is shown in the last three columns of first row. We can see that increasing \( \gamma \) deteriorates the quality of the barycenter which can have a negative impact on barycentric accuracy. Similarly, the effect of changing \( \theta \)
TABLE IV
Performance against attacks on CIFAR-10. $\epsilon = 0.5$ for L2PGD and L2CarliniWagner. $\epsilon = 8/255$ for Linfpgd, FGSM, Linfdeepfool.

<table>
<thead>
<tr>
<th>Attack</th>
<th>AA (in %)</th>
<th>BA (in %)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2PGD [7]</td>
<td>2.65</td>
<td>32.51</td>
</tr>
<tr>
<td>L2CarliniWagner [13]</td>
<td>0.06</td>
<td>83.07</td>
</tr>
<tr>
<td>Linfpgd [7]</td>
<td>0</td>
<td>14.62</td>
</tr>
<tr>
<td>FGSM [5]</td>
<td>37.48</td>
<td>43.11</td>
</tr>
<tr>
<td>Linfdeepfool [20]</td>
<td>1.81</td>
<td>78.11</td>
</tr>
<tr>
<td>AutoAttack [2]</td>
<td>0</td>
<td>10.15</td>
</tr>
</tbody>
</table>

and $\tau$ is given in second and third row respectively. We observe that decreasing $\theta$ leads to blurring while increasing $\theta$ generates reasonable quality of barycenter. Further, a large $\tau$ also deteriorates the quality of barycenter.

![Fig. 1. First row: $\gamma = .02, .01, 1, 5$ with $\theta = 1e5, \tau = 0.2$. Second row: $\gamma=0.02, \theta = 1e3, 1e4, 1e6, 1e7$ and $\tau = 0.2$. Last row: $\gamma=0.02, \theta = 1e5$ and $\tau = 0.005, 0.05, 1, 5$.](image)

D. Subjective Evaluation
We show the barycenters in Figure 2. In the figure we can observe that the barycenters obtained under $B6$ setting are close to the respective original images. In Figure 3, we show the gradcam heatmaps. We can observe that the heatmaps for barycenters are closer to the heatmaps of respective original images compared to that of attacked images. Especially, in case of MNIST and Fashion-MNIST, we can clearly observe that heatmaps of the clean images and respective barycenters show a high degree of similarity. In case of CIFAR-10, the heatmap of attacked image is very different from the heatmap of clean image indicating that the model may not be able to correctly classify the attacked image. Thus, empirically we can observe that the barycenter can mitigate the $L_\infty$ attacks effectively without any training.

![Fig. 2. Clean image, Barycenter ($B6$) for MNIST, CIFAR and Fashion-MNIST.](image)

![Fig. 3. Gradcam heatmaps of clean images, $L_\infty$ attacked images and Barycenters of MNIST, CIFAR-10 and Fashion-MNIST.](image)

V. Conclusion
We present a novel adversarial defense technique and demonstrate that models can be defended without the need of any adversarial training. We formulate a Wasserstein barycenter objective and analytically solve for the barycenter. In order to obtain the barycenter, we use the given image and derive its marginals using rotation and translation. Our empirical analysis using GradCam show that the barycenters possess the traits of clean images and thus have a better probability of being correctly classified. We perform elaborate experiments to qualitatively and quantitatively show that barycenters exhibit a strong defense against a variety of adversarial attacks.
REFERENCES


